

面向多源出行数据的可视化查询模型（偏向综述类的论文，可以提出一种设想、模型、框架等）

1、引言

世界上大约有一半的人生活在城市，每天产生 PB 级出行数据，如公共交通记录、车流量监控记录、出租车记录、智能手机记录、社交网络签到记录等。出行数据不仅成功应用于城市计算的基本问题，如城市规划[1]、交通状况[2]、空气质量[3]、灾难分析[4]等，而且在城市内人群运动规律上得出重要结论[5, 6]。

出行数据具有时空特性，而且体量大、维度高，但是与常见的轨迹数据不同的是出行数据的类型更多样而且时空粒度不一致。受出行数据特点的限制，已有的分析方法往往限定在特定类型数据，而且分析目标单一、分析周期较长[7]。

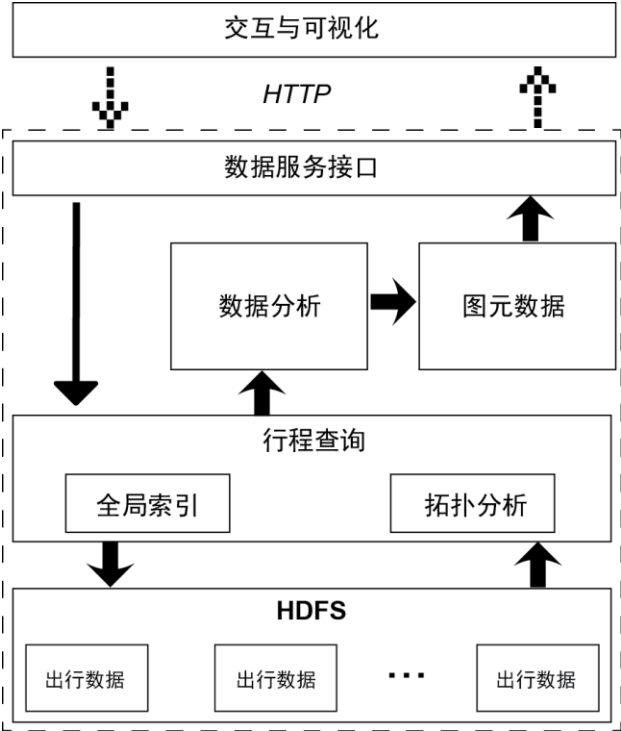
可视分析方法提供图形化交互界面，允许用户探索式分析大规模异构数据，因而可视分析方法对应用系统提出更高的要求。面对多源出行数据，可视化分析至少需要解决三方面的困难：1）如何抽象出行数据使得多源异构数据具有相同的结构；2）如何组织使得交互探索具有实时性；3）如何可视化表达多源数据使其和谐地呈现出行模式。

本文提出了从数据组织、存储到索引、查询，最后到交互式查询、可视化和分析的集成模型。模型包含多个步骤，首先分析出行数据的基本结构，抽象出统一的出行数据表达形式；在此基础上，数据按时间和空间分类存储于 HDFS 文件系统上；为提高数据探索的效率，构建了基于哈希的双向链式索引；本文提出了基于空间拓扑的查询模式，查询算法基于 Spark 实现；为方便数据探索，本文设计了基于 sketch 的交互界面；同时本文提供了多种可视化设计用于不同的分析目标。利用本文的可视化查询与分析系统，一些重要的出行模式都能很容易地被发现。

2、相关工作

- 基于统计的数据分析方法、数据挖掘分析方法、可视化分析方法
- 城市数据存储和管理
- 城市数据可视化技术

3、系统总体结构



本文的可视化查询与分析系统总体分为两部分，其中交互与可视化模块负责传递用户意图并且反馈分析结果供其理解和决策，而与之对应的数据服务模块主要是获取期望的数据并返回可以理解的数据图元。两部分之间经 HTTP 协议连接传递少量数据，降低数据规模和复杂性对用户分析数据的影响。数据服务模块中包含数据存储、轨迹查询、数据分析和可视映射四项功能。出行数据按时间和空间分布存储于 HDFS 的不同数据节点上。轨迹查询则将用户意图转换成查询条件从全局索引中寻找适合的出行轨迹，并计算轨迹与查询空间约束的拓扑关系。查询基于 spark 模型，在不同数据节点上分布计算，并在统一内存中交换数据。数据分析主要实现分类、聚类、统计等分析计算，之后将计算结果映射为特定的图元数据，并通过数据服务接口返回至交互与可视化模块。（参考其他可视化论文对框架的描述，感觉文字有些少）

4、可视化查询

可视化查询的溯源

可视化查询的目标

可视化查询需要哪些操作，与普通的查询有何不同

5、数据抽象和数据存储

5.1 数据抽象

尽管记录出行数据的设备各有不同，但是数据中一般都会包含出行的对象（人或车）、运动对象的位置、记录的时间以及一些其他属性，如车辆的状态、速度或进出站标识等。如果仅考虑这些数据的共性，那么多源出行数据模式可以统一定义为四元组的集合，即

$$\mathcal{M} = \langle O, L, \Gamma, \Omega \rangle$$

其中， O 为出行对象集合， L 为地理位置， Γ 表示记录时间， Ω 为其他各类属性集合。

给定一个时间段，任意对象必将产生按照数据模式 \mathcal{M} 的出行数据集合，称为行程。一般地，行程可以定义为

$$\gamma = \{(o_i, l_1, t_1, \omega_1), (o_i, l_2, t_2, \omega_2), \dots (o_i, l_n, t_n, \omega_n) \mid i, n \in \mathbb{N}\}$$

而行程中每个记录则称为行程点，每个行程由时间连续的行程点集合组成。实际行程中，出行对象可以处于运动状态也可能处于停留状态，尽管本文主要研究前者但数据模型并不影响同时对上述两种状态的研究。处于运动状态的对象，在行程上通常表现为相邻行程点位置不同，即

$$\gamma_{i \cdot l_k} \neq \gamma_{i \cdot l_{k+1}}$$

受数据记录方式的影响，不同来源的行程在时间具有不同的数据粒度，有些时间间隔比较均匀，而更多的时间间隔差异较大。比如，浮动车数据时间均匀间隔，通常几十秒到几十分钟记录一次；手机通信数据则时间跨度大，有的行程点间隔几十秒而有的则间隔数小时。另一方面，受出行交通工具和时间不均衡的影响，行程空间上跨度也极为不均衡。连续的行程点有的相隔几米，而有的则达到数公里。数据时空分布不均衡直接导致行程存储上的困难。

5.2 数据存储

为保证行程访问的效率，数据需要进行合理的划分。划分的基本原则是：相关性较强的数据物理存储上相邻【查文献】。从行程的数据特征来看，可以有如下四种划分方法：

- 1) 按时间分段存储，即按一定的时间段（小时、天等）将行程分割，不同时间段内的行程存储于不同数据块上。时间段划分非常均匀、而且操作简单，更重要的是方便按时间获取行程。对于时间间隔均匀的行程，如浮动车按时间分段存储较好。然而对于时间上分布不均衡的行程，尽管划分的时间间隔相同，但是其数据量可能差别很大。
- 2) 按空间分块存储，即在地理上将数据置于不同的区块内，每个区块内存储经过此地的行程点。由于城市内区域功能不尽相同，均匀的区域划分会导致不同区域数据不均衡。层次划分（四叉树、R 树等）能调节区域内的数据均衡性，是常用的划分方法[seti]。在某

些情况下，当数据本身就具备一定的区域特性时，直接利用这种特性划分效果更好。例如，手机通信数据中，基站的设置本身就是保证用户通信的均衡性。因此按基站覆盖范围划分数据较为合理。

- 3) 按出行对象分块存储，即每个出行对象的行程单独存储于独立的数据块。由于城市中人群的性质不同，其行程尺度差异较大。对于上班族而言，其出行点基本固定在几个位置；而出租车司机则行程点多且分布在城市各地。简单的按出行对象存储会造成出行尺度小的数据块空间浪费，适当的数据合并是减少空间浪费的主要方式。
- 4) 以上三种方式的混合划分。例如，对于尺度大的行程，先按空间划分再按时间划分，或者先根据出行对象划分，当出行数据集很大时再按照时间划分等。本文的系统中，出行数据主要采用混合划分的方式，确保每个数据块中的出行数据尽量均衡，而且读取性能有所保证。

按照上述规则划分的数据块可以认为是时间或空间上具有关联性的行程集合。因此，可以通过时间、空间值或出行对象等属性唯一性地标识数据块。形式上，数据块是上述三个属性的唯一映射，即：

$$B_i = f(o, t, l)$$

本文采取的一种存储方式为每个数据块记录一段时间内经过特定地理区域内的所有行程点。

6、数据索引与检索

彼此之间相互关联，建立双向哈希索引

基于 spark 实现多种拓扑的查询：enter, cross, pass, cover

6.1 双向链接哈希索引

可视化查询一般会指定特定的时间和空间范围，范围具有较大的可伸缩性，可以是很大的区域或时间段，也可以是很小的区域或时间段。受查询范围的限制，查询得到的行程往往只是局部。为了提高查询的效率，设计适当的索引是非常必要的。研究人员针对时空数据设计了诸多索引，但索引大都基于较大的时间和空间划分，难以在任意指定的时空范围内快速获取行程路径。



考虑到行程的时间连续性，行程上的点可以依次按照所在空间区域连接。此外，由于同一空间区域内的行程点非常多，为提高查询效率可以在每个空间区域内设立键值对桶，通过哈希函数获取指定出行对象特定时间段的行程。

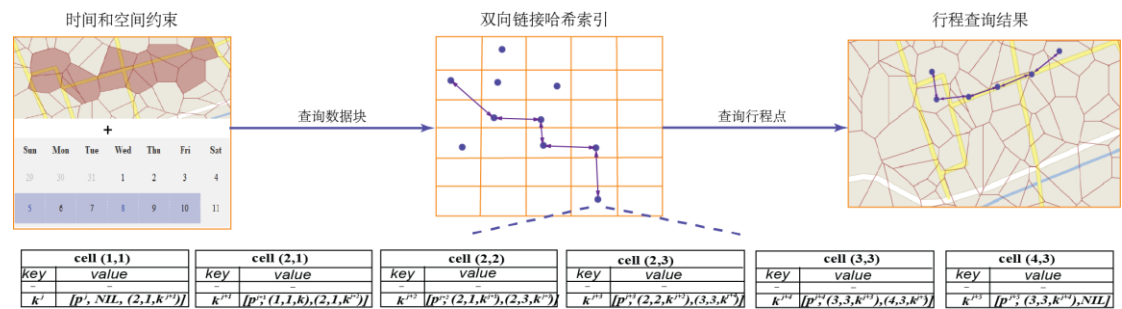
构建索引的过程（图 2）可分为三个阶段：地图划分、存储行程点和连接行程。在地图划分阶段，根据数据特点将地图分成适当的区域集合，使得经过每个区域的行程点数量尽量平衡，这些区域可以是均匀大小的网格也可以是非均匀的。划分地图后，每个区域都建立一个键值对桶，经过该区域的行程点都将其行程对象及时间戳插入到桶内。最后，相同行程对象按照时间先后顺序双向链接，将各个键值对桶相关串联。

图 3 描述了双向链接的行程索引结构，总体上类似于 Grid-File[ref]具有两层索引，不同的是本文的索引要求在数据空间上进行均匀划分，而不是时间或空间上均匀划分。第一级索引定义在时间和空间维度，用于获取对应的键值对桶。第二级索引定义在时间和行程对象维度，用于在键值对桶中唯一地获取行程点数据。

如前所述，本文的数据块可以由时间、空间标识，对应于第一级索引。而数据块内的行程点位置则由第二级索引负责。实现上，键值对桶中的主键是时间戳和行程对象的函数，而

值为数据块内的物理偏移以及前后行程点所在数据块的标识。实际查询中，可以一次读取整个数据块为下次查询提供缓存。

6.2 行程查询



7、轨迹数据可视化与交互式分析

提供基于力引导的边绑定展示轨迹的空间特征、用 **stack graph** 展示轨迹的时间变化特征。

8、结论

参考文献

[1] Yu Zheng, Yanchi Liu, Jing Yuan, Xing Xie, Urban computing with taxicabs, UbiComp 2011

[2] Visual Traffic Jam Analysis Based on Trajectory Data, *IEEE Transactions on Visualization and Computer Graphics (VAST'13)*, 19(12):2159-2168, 2013.

[3] Yu Zheng, Furui Liu, Hsun-Ping Hsie. U-Air: when urban air quality inference meets big data. KDD 2013

[4] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, R. Shibasaki. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. KDD 2013

[5] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility, *Science* 19 February 2010: 327 (5968), 1018-1021

[6] Marta C. González, Cesar A. Hidalgo R., and Albert-László Barabási. Understanding individual human mobility patterns

[7] Desheng zhang

Exploring human mobility with multi-source data at extremely large metropolitan scales.